k-MLE: A fast algorithm for learning statistical mixture models*

Frank Nielsen
Sony Computer Science Laboratories, Inc
3-14-13 Higashi Gotanda
141-0022 Shinagawa-Ku, Tokyo, Japan

E-mail:Frank.Nielsen@acm.org

June 2011 (revised March 2012)

Abstract

We describe k-MLE, a fast and efficient local search algorithm for learning finite statistical mixtures of exponential families such as Gaussian mixture models. Mixture models are traditionally learned using the expectation-maximization (EM) soft clustering technique that monotonically increases the incomplete (expected complete) likelihood. Given prescribed mixture weights, the hard clustering k-MLE algorithm iteratively assigns data to the most likely weighted component and update the component models using Maximum Likelihood Estimators (MLEs). Using the duality between exponential families and Bregman divergences, we prove that the local convergence of the complete likelihood of k-MLE follows directly from the convergence of a dual additively weighted Bregman hard clustering. The inner loop of k-MLE can be implemented using any k-means heuristic like the celebrated Lloyd's batched or Hartigan's greedy swap updates. We then show how to update the mixture weights by minimizing a cross-entropy criterion that implies to update weights by taking the relative proportion of cluster points, and reiterate the mixture parameter update and mixture weight update processes until convergence. Hard EM is interpreted as a special case of k-MLE when both the component update and the weight update are performed successively in the inner loop. To initialize k-MLE, we propose k-MLE++, a careful initialization of k-MLE guaranteeing probabilistically a global bound on the best possible complete likelihood.

Keywords: exponential families, mixtures, Bregman divergences, expectation-maximization (EM), k-means loss function, Lloyd's k-means, Hartigan and Wong's k-means, hard EM, sparse EM.

1 Introduction

1.1 Statistical mixture models

A statistical mixture model [34] $M \sim m$ with $k \in \mathbb{N}$ weighted components has underlying probability distribution:

*Research performed during the January-June 2011 period. A preliminary shorter version appeared in IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2012.

$$m(x|w,\theta) = \sum_{i=1}^{k} w_i p(x|\theta_i), \tag{1}$$

with $w = (w_1, ..., w_k)$ and $\theta = (\theta_1, ..., \theta_k)$ denoting the mixture parameters: The w_i's are positive weights summing up to one, and the θ_i's denote the individual component parameters. (Appendix E summarizes the notations used throughout the paper.)

Mixture models of d-dimensional Gaussians[1] are the most often used statistical mixtures [34]. In that case, each component distribution $N(\mu_i, \Sigma_i)$ is parameterized by a mean vector $\mu_i \in \mathbb{R}^d$ and a covariance matrix $\Sigma_i \succ 0$ that is symmetric and positive definite. That is, $\theta_i = (\mu_i, \Sigma_i)$. The Gaussian distribution has the following probability density defined on the support $\mathbb{X} = \mathbb{R}^d$:

$$p(x; \mu_i, \Sigma_i) = \frac{1}{(2\pi)^{\frac{d}{2}} \sqrt{|\Sigma_i|}} e^{-\frac{1}{2} M_{\Sigma_i^{-1}}(x-\mu_i, x-\mu_i)}, \tag{2}$$

where M_Q denotes the squared Mahalanobis distance [12]

$$M_Q(x,y) = (x-y)^T Q (x-y), \tag{3}$$

defined for a symmetric positive definite matrix $Q \succ 0$ ($Q_i = \Sigma_i^{-1}$, the precision matrix).

To draw a random variate from a Gaussian mixture model (GMM) with k components, we first draw a multinomial variate $z \in \{1, ..., k\}$, and then sample a Gaussian variate from $N(\mu_z, \Sigma_z)$. A multivariate normal variate x is drawn from the chosen component $N(\mu, \Sigma)$ as follows: First, we consider the Cholesky decomposition of the covariance matrix: $\Sigma = CC^T$, and take a d-dimensional vector with coordinates being random standard normal variates: $y = [y_1 \ ... \ y_d]^T$ with $y_i = \sqrt{-2 \log u_1} \cos(2\pi u_2)$ (for u_1 and u_2 uniform random variates in $[0, 1)$). Finally, we assemble the Gaussian variate x as $x = \mu + Cy$. This drawing process emphasizes that sampling a statistical mixture is a *doubly stochastic process* by essence: First, we sample a multinomial law for choosing the component, and then we sample the variate from the selected component.

Figure 1(b) shows a GMM with $k = 32$ components learned from a color image modeled as a 5D xyRGB point set (Figure 1(a)). Since a GMM is a *generative model*, we can sample the GMM to create a "sample image" as shown in Figure 1(c). Observe that low frequency information of the image is nicely modeled by GMMs. Figure 2(f) shows a GMM with $k = 32$ components learned from a color image modeled as a high-dimensional point set. Each $s \times s$ color image patch anchored at (x, y) is modeled as a point in dimension $d = 2 + 3s^2$. GMM representations of images and videos [21] provide a compact feature representation that can be used in many applications, like in information retrieval (IR) engines [14].

In this paper, we consider the general case of mixtures of distributions belonging the same exponential family [50], like Gaussian mixture models [24] (GMMs), Rayleigh mixture models [47] (RMMs), Laplacian mixture models (LMMs)[4], Bernoulli mixture models [5] (BMMs), Multinomial Mixture models [46] (MMMs), Poisson Mixture Models (PMMs) [28], Weibull Mixture Models [15] (WeiMMs), Wishart Mixture Models [22] (WisMM), etc.

1.2 Contributions and prior work

Expectation-Maximization [18] (EM) is a traditional algorithm for learning finite mixtures [34]. Banerjee et al. [9] proved that EM for mixture of exponential families amounts to perform equiv-

[1] Also called MultiVariate Normals (MVNs) in software packages.

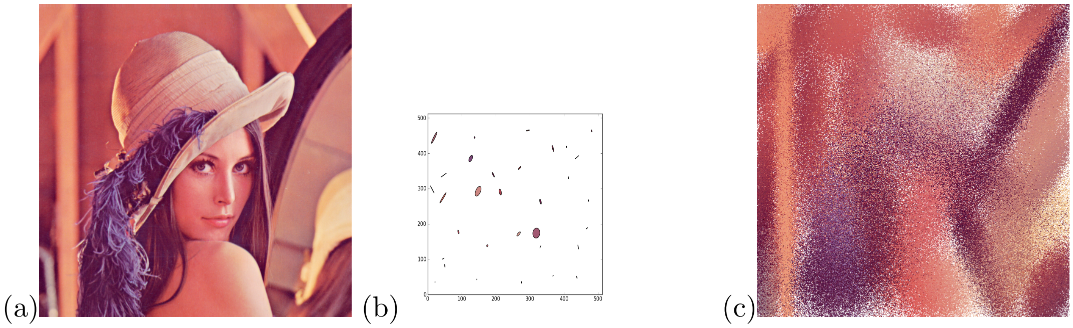
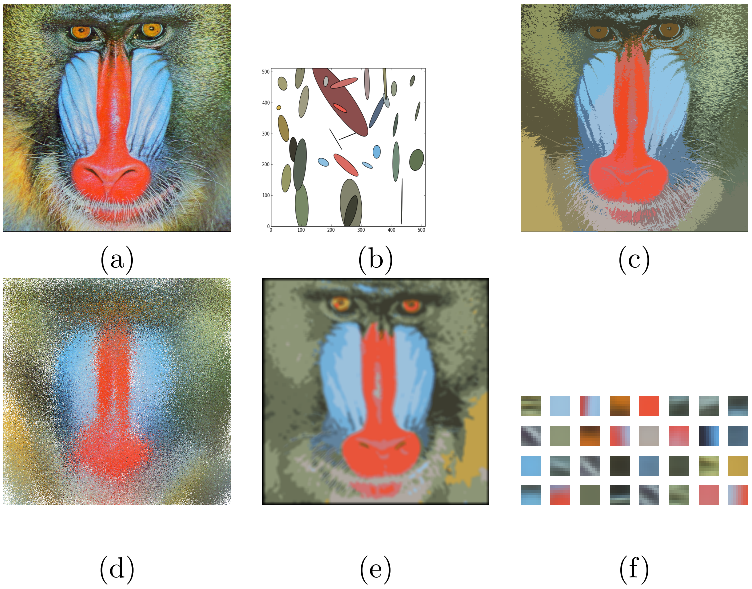

(a) (b) (c)

Figure 1: A RGB color image (a) is interpreted as a 5D xyRGB point set on which a Gaussian mixture model (GMM) with $k = 32$ components is trained (b). Drawing many random variates from the generative GMM yields a sample image(c) that keeps low-frequency visual information.


(a) (b) (c)

(d) (e) (f)

Figure 2: Modeling a color image using a Gaussian mixture model (GMM): (a) Image `Baboon` source image, (b) a 5D 32-GMM modeling depicted by its covariance ellipses, (c) hard segmentation using the GMM, (d) sampling the 5D GMM, (e) Mean colors (8×8 patches) for GMM with patch size $s = 8$, (f) patch mean μ for $s = 8$ patch size width.

alently a soft Bregman clustering. Furthermore, this EM-Bregman soft clustering equivalence was extended to total Bregman soft clustering for curved exponential families [29]. Although mathematically convenient, we should remember that mixture data should be hard clustered as each observation should emanate from exactly one component.

It is well-known that k-means clustering technique can be interpreted as a limit case of EM for isotropic Gaussian mixtures [37]. Kearns et al. [26] casted further light on the hard/soft relationship using an information-theoretic analysis of hard k-means and soft expectation-mazimization assignments in clustering. Banerjee et al [7] proved a mathematical equivalence between the estimation of maximum likelihood of exponential family mixtures (MLME, Maximum Likelihood Mixture Estimation) and a rate distortion problem for Bregman divergences. Furthermore, Banerjee et al. [8] proposed the hardened expectation for the special case of von Mises-Fisher mixtures (hard EM, Section 4.2 of [8]) for computational efficiency.

In this paper, we build on the duality between Bregman divergences and exponential families [9] to design k-MLE that iteratively (1) assigns data to mixture components, (2) update mixture parameters à la k-means and repeat step (1) until local convergence, (3) update weights and reiterate from (1) until local convergence (see Algorithm 1). We prove that k-MLE maximizes monotonically the complete likelihood function. We also discuss several initialization strategies and describe a probabilistic initialization k-MLE++ with guaranteed performance bounds.

The paper is organized as follows: Section 2 recall the basic notions of exponential families, Legendre transform, Bregman divergences, and demonstrate the duality between Bregman divergences and exponential families to study the Maximum Likelihood Estimator (MLE). Section 3 presents the framework of k-MLE for mixtures with prescribed weights, based on the Bregman-exponential family duality. The generic k-MLE algorithm is described in Section 4, and Section 5 discusses on proximity location data-structures to speed up the assignment step of the algorithm. Section 6 presents k-MLE++, a probabilistic initialization of k-MLE. Finally, Section 7 concludes the paper and discusses on avenues for future research.

2 Preliminaries

2.1 Exponential family

An exponential family [13] E_F is a set of parametric probability distributions

$$E_F = \{p_F(x;\theta) \mid \theta \in \Theta\} \tag{4}$$

whose probability density[2] can be decomposed canonically as

$$p_F(x;\theta) = e^{\langle t(x),\theta\rangle - F(\theta) + k(x)} \tag{5}$$

where $t(x)$ denotes the sufficient statistics, θ the natural parameter, $F(\theta)$ the log-normalizer, and $k(x)$ a term related to an optional auxiliary carrier measure. $\langle x, y\rangle$ denotes the inner product (i.e., $x^T y$ for vectors $\mathrm{tr}(X^T Y)$ for matrices, etc.). Let

$$\Theta = \left\{\theta \mid \int p_F(x;\theta)\mathrm{d}x < \infty\right\} \tag{6}$$

[2] For sake of simplicity and brevity, we consider without loss of generality in the remainder continuous random variables on $\mathbb{R}^d$. We do not introduce the framework of probability measures nor Radon-Nikodym densities.

denotes the natural parameter space. The dimension D of the natural parameter space is called the order of the family. For the d-variate Gaussian distribution, the order is $D = d + \frac{d(d+1)}{2} = \frac{d(d+3)}{2}$. It can be proved using the Cauchy-Schwarz inequality [13] that the log-normalizer[3] F is a strictly convex and differentiable function on an open convex set Θ. The log-density of an exponential family is

$$l_F(x;\theta) = \langle t(x), \theta\rangle - F(\theta) + k(x) \tag{7}$$

To build an exponential family, we need to choose a basic density measure on a support $\mathcal{X}$, a sufficient statistic $t(x)$, and an auxiliary carrier measure term $k(x)$. Taking the log-Laplace transform, we get

$$F(\theta) = \int_{x\in\mathbb{X}} e^{\langle t(x),\theta\rangle + k(x)} \mathrm{d}x, \tag{8}$$

and define the natural parameter space as the θ values ensuring convergence of the integral.

In fact, many usual statistical distributions such as the Gaussian, Gamma, Beta, Dirichlet, Poisson, multinomial, Bernoulli, von Mises-Fisher, Wishart, Weibull are exponential families in disguise. In that case, we start from their probability density or mass function to retrieve the canonical decomposition of Eq. 5. See [36] for usual canonical decomposition examples of some distributions that includes a bijective conversion function $\theta(\lambda)$ for going from the usual λ-parameterization of the distribution to the θ-parametrization.

Furthermore, exponential families can be parameterized canonically either using the natural coordinate system θ, or by using the dual moment parameterization η (also called mean value parameterization) arising from the Legendre transform (see Appendix B for the case of Gaussians).

2.2 Legendre duality and convex conjugates

For a strictly convex and differentiable function $F : \mathbb{N} \to \mathbb{R}$, we define its convex conjugate by

$$F^*(\eta) = \sup_{\theta\in\mathbb{N}} \{\underbrace{\langle \eta, \theta\rangle - F(\theta)}_{l_F(\eta;\theta)}\} \tag{9}$$

The maximum is obtained for $\eta = \nabla F(\theta)$ and is unique since F is convex $\nabla^2_\theta l_F(\eta;\theta) = -\nabla^2 F(\theta) \prec 0$:

$$\nabla_\theta l_F(\eta;\theta) = \eta - \nabla F(\theta) = 0 \Rightarrow \eta = \nabla F(\theta) \tag{10}$$

Thus strictly convex and differentiable functions come in pairs (F, F^*) with gradients being functional inverses of each other $\nabla F = (\nabla F^*)^{-1}$ and $\nabla F^* = (\nabla F)^{-1}$. Legendre transform is an involution: $(F^*)^* = F$ for strictly convex and differentiable functions. In order to compute F^*, we only need to find the functional inverse $(\nabla F)^{-1}$ of ∇F since

$$F^*(\eta) = \langle (\nabla F)^{-1}(\eta), \eta\rangle - F((\nabla F)^{-1}(\eta)). \tag{11}$$

However, this inversion may require numerical solving when no analytical expression of ∇F^{-1} is available. See for example the gradient of the log-normalizer of the Gamma distribution [36], the Dirichlet or von Mises-Fisher distributions [8].

[3] Also called in the literature as the log-partition function, the cumulant function, or the log-Laplace function.

2.3 Bregman divergence

A Bregman divergence B_F is defined for a strictly convex and differentiable generator F as

$$B_F(\theta_1 : \theta_2) = F(\theta_1) - F(\theta_2) - \langle \theta_1 - \theta_2, \nabla F(\theta_2)\rangle. \tag{12}$$

The Kullback-Leibler divergence (relative entropy) between two members $p_1 = p_F(x;\theta_1)$ and $p_2 = p_F(x;\theta_2)$ of the same exponential family amounts to compute a Bregman divergence on the corresponding swapped natural parameters:

$$\begin{aligned}
\mathrm{KL}(p_1 : p_2) &= \int_{x\in\mathbb{X}} p_1(x)\log\frac{p_1(x)}{p_2(x)}\mathrm{d}x, && (13)\\
&= B_F(\theta_2 : \theta_1), && (14)\\
&= F(\theta_2) - F(\theta_1) - \langle \theta_2 - \theta_1, \nabla F(\theta_1)\rangle && (15)
\end{aligned}$$

The proof follows from the fact that $E[t(X)] = \int_{x\in\mathbb{X}} t(x)p_F(x;\theta)\mathrm{d}x = \nabla F(\theta)$ [39]. Using Legendre transform, we further have the following equivalences of the relative entropy:

$$\begin{aligned}
B_F(\theta_2 : \theta_1) &= B_{F*}(\eta_1 : \eta_2), && (16)\\
&= \underbrace{F(\theta_2) + F^*(\eta_1) - \langle \theta_2, \eta_1\rangle}_{C_F(\theta_2:\eta_1) = C_{F^*}(\eta_1:\theta_2)}, && (17)
\end{aligned}$$

where $\eta = \nabla F(\theta)$ is the dual moment parameter (and $\theta = \nabla F^*(\eta)$). Information geometry [3] often considers the canonical divergence C_F of Eq. 17 that uses the mixed coordinate systems θ/η, while computational geometry [12] tends to consider dual Bregman divergences, B_F or B_{F^*}, and visualize structures in one of those two canonical coordinate systems. Those canonical coordinate systems are dually orthogonal since $\nabla^2 F(\theta)\nabla^2 F^*(\eta) = I$, the identity matrix.

2.4 Maximum Likelihood Estimator (MLE)

For exponential family mixtures with a single component $M \sim E_F(\theta_1)$ ($k = 1$, $w_1 = 1$), we easily estimate the parameter θ_1. Given n independent and identically distributed observations $x_1, ..., x_n$, the Maximum Likelihood Estimator (MLE) is maximizing the likelihood function:

$$\begin{aligned}
\hat{\theta} &= \mathrm{argmax}_{\theta\in\Theta} L(\theta; x_1, ..., x_n), && (18)\\
&= \mathrm{argmax}_{\theta\in\Theta} \prod_{i=1}^{n} p_F(x_i;\theta), && (19)\\
&= \mathrm{argmax}_{\theta\in\Theta} e^{\sum_{i=1}^{n}\langle t(x_i),\theta\rangle - F(\theta) + k(x_i)} && (20)
\end{aligned}$$

For exponential families, the MLE reports a unique maximum since the Hessian of F is positive definite ($X \sim E_F(\theta) \Rightarrow \nabla^2 F = \mathrm{var}[t(X)] \succ 0$):

$$\nabla F(\hat{\theta}) = \frac{1}{n}\sum_{i=1}^{n} t(x_i) \tag{21}$$

Exponential Family	⇔	Dual Bregman divergence
$p_F(x\mid\theta)$		B_{F^*}
Spherical Gaussian	⇔	Squared Euclidean divergence
Multinomial	⇔	Kullback-Leibler divergence
Poisson	⇔	I-divergence
Geometric	⇔	Itakura-Saito divergence
Wishart	⇔	log-det/Burg matrix divergence

Table 1: Some examples illustrating the duality between exponential families and Bregman divergences.

The MLE is consistent and efficient with asymptotic normal distribution:

$$\hat{\theta} \sim N\left(\theta, \frac{1}{n} I_F^{-1}(\theta)\right), \tag{22}$$

where I_F denotes the Fisher information matrix:

$$I_F(\theta) = \mathrm{var}[t(X)] = \nabla^2 F(\theta) = (\nabla^2 G(\eta))^{-1} \tag{23}$$

(This proves the convexity of F since the covariance matrix is necessarily positive definite.) Note that the MLE may be biased (for example, normal distributions).

By using the Legendre transform, the log-density of an exponential family can be interpreted as a Bregman divergence [9]:

$$\log p_F(x;\theta) = -B_{F^*}(t(x):\eta) + F^*(t(x)) + k(x) \tag{24}$$

Table 1 reports some illustrating examples of the Bregman divergence ↔ exponential family duality. Let us use the Bregman divergence-exponential family duality to prove that

$$\hat{\theta} = \arg\max_{\theta\in\Theta} \prod_{i=1}^{n} p_F(x_i;\theta) = \nabla F^{-1}\left(\sum_{i=1}^{n} t(x_i)\right). \tag{25}$$

Maximizing the average log-likelihood $\bar{l} = \frac{1}{n}\log L$, we have:

$$\begin{aligned}
\max\nolimits_{\theta\in\mathbb{N}} \quad & \bar{l}(\theta; x_1, ..., x_n) = \frac{1}{n}\sum_{i=1}^{n}(\langle t(x_i), \theta\rangle - F(\theta) + k(x_i)) && (26)\\
\max\nolimits_{\theta\in\mathbb{N}} \quad & \frac{1}{n}\sum_{i=1}^{n} -B_{F^*}(t(x_i):\eta) + F^*(t(x_i)) + k(x_i) && (27)\\
\equiv \min\nolimits_{\eta\in\mathbb{M}} \quad & \frac{1}{n}\sum_{i=1}^{n} B_{F^*}(t(x_i):\eta) && (28)
\end{aligned}$$

Since right-sided Bregman centroids defined as the minimum average divergence minimizers coincide always with the center of mass [9] (independent of the generator F), it follows that

$$\hat{\eta} = \frac{1}{n}\sum_{i=1}^{n} t(x_i) = \nabla F(\hat{\theta}). \tag{29}$$

It follows that $\hat{\eta} = (\nabla F)^{-1}(\frac{1}{n}\sum_{i=1}^n t(x_i))$.

In information geometry [3], the point $\hat{P}$ with η-coordinate $\hat{\eta}$ (and θ-coordinate $\nabla F^{-1}(\hat{\eta}) = \hat{\theta}$) is called the *observed* point. The best average log-likelihood reached by the MLE at $\hat{\eta}$ is

$$\begin{aligned}
l(\hat{\theta}; x_1, ..., x_n) &= \frac{1}{n}\sum_{i=1}^n (-B_{F^*}(t(x_i) : \hat{\eta}) + F^*(t(x_i)) + k(x_i)), & (30)\\
&= \frac{1}{n}\sum_{i=1}^n (-F^*(t(x_i)) + F^*(\hat{\eta}) + \langle t(x_i) - \hat{\eta}, \nabla F^*(\hat{\eta})\rangle + F^*(t(x_i)) + k(x_i)), & (31)\\
&= F^*(\hat{\eta}) + \frac{1}{n}\sum_{i=1}^n k(x_i) + \left\langle \underbrace{\frac{1}{n}\sum_{i=1}^n t(x_i) - \hat{\eta}}_{0}, \hat{\theta} \right\rangle, & (32)\\
&= F^*(\hat{\eta}) + \frac{1}{n}\sum_{i=1}^n k(x_i). & (33)
\end{aligned}$$

The Shannon entropy $H_F(\theta)$ of $p_F(x;\theta)$ is $H_F(\theta) = -F^*(\eta) - \int k(x) p_F(x;\theta)\mathrm{d}x$ [39]. Thus the maximal likelihood is related to the minimum entropy (i.e., reducing the uncertainty) of the empirical distribution.

Another proof follows from the Appendix A where it is recalled that the Bregman information [9] (minimum of average right-centered Bregman divergence) obtained for the center of mass is a Jensen diversity index. Thus we have

$$\begin{aligned}
\bar{l} &= -J_{F^*}(\sum_{i=1}^n t(x_i)) + \frac{1}{n}\sum_{i=1}^n F^*(t(x_i)) + \frac{1}{n}\sum_{i=1}^n k(x_i), & (34)\\
&= -\left(\sum_{i=1}^n F^*(t(x_i)) - F^*(\hat{\eta})\right) + \frac{1}{n}\sum_{i=1}^n F^*(t(x_i)) + \frac{1}{n}\sum_{i=1}^n k(x_i), & (35)\\
&= F^*(\hat{\eta}) + \frac{1}{n}\sum_{i=1}^n k(x_i) & (36)
\end{aligned}$$

Appendix B reports the dual canonical parameterizations of the multivariate Gaussian distribution family.

3 k-MLE: Learning mixtures with given prescribed weights

Let $\mathcal{X} = \{x_1, ..., x_n\}$ be a sample set of independently and identically distributed observations from a finite mixture $m(x|w,\theta)$ with k components. The joint probability distribution of the observed observations x_i's with the missing component labels z_i's is

$$p(x_1, z_1, ..., x_n, z_n | w, \theta) = \prod_{i=1}^n p(z_i|w) p(x_i|z_i, \theta) \quad (37)$$

To optimize the joint distribution, we could test (theoretically) all the k^n labels, and choose the best assignment. This is not tractable in practice since it is exponential in n for $k > 1$. Since we do not observe the latent variables $z_1, ..., z_n$, we marginalize the hidden variables to get

$$p(x_1, ..., x_n|w, \theta) = \prod_{i=1}^{n} \sum_{j=1}^{k} p(z_i = j|w) p(x_i|z_i = j, \theta_j) \tag{38}$$

The average log-likelihood function is

$$\begin{aligned} \bar{l}(x_1, ..., x_n|w, \theta) &= \frac{1}{n} \log p(x_1, ..., x_n|w, \theta), &(39)\\ &= \frac{1}{n} \sum_{i=1}^{n} \log \sum_{j=1}^{k} p(z_i = j|w) p(x_i|z_i = j, \theta_j). &(40) \end{aligned}$$

Let $\delta_j(z_i) = 1$ if and only if x_i has been sampled from the jth component, and 0 otherwise. We have the complete average log-likelihood that is mathematically rewritten as

$$\begin{aligned} \bar{l}(x_1, z_1, ..., x_n, z_n|w, \theta) &= \frac{1}{n} \sum_{i=1}^{n} \log \prod_{j=1}^{k} (w_j p_F(x_i|\theta_j))^{\delta_j(z_i)} &(41)\\ &= \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{k} \delta_j(z_i)(\log p_F(x_i|\theta_j) + \log w_j) &(42) \end{aligned}$$

Using the bijection between exponential families and dual Bregman divergences [9], we have the mathematical equivalence $\log p_F(x|\theta_j) = -B_{F^*}(t(x) : \eta_j) + F^*(t(x)) + k(x)$, where $\eta_j = \nabla F(\theta_j)$ is the moment parameterization of the j-th component exponential family distribution. It follows that the complete average log-likelihood function is written as

$$\begin{aligned} \bar{l}(x_1, ..., x_n|w, \theta) &= \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{k} \delta_j(z_i)(-B_{F^*}(t(x_i) : \eta_j) + F^*(t(x_i)) + k(x_i) + \log w_j) &(43)\\ &= \left(\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{k} \delta_j(z_i)(-B_{F^*}(t(x_i) : \eta_j) + \log w_j) \right) + \frac{1}{n} \sum_{i=1}^{n} F^*(t(x_i)) + k(x_i) &(44) \end{aligned}$$

By removing the constant terms $\frac{1}{n} \sum_{i=1}^{n} (F^*(t(x_i)) + k(x_i))$ independent of the mixture moment parameters (the η's), maximizing the complete average log-likelihood amounts to equivalently minimize the following loss function:

$$\begin{aligned} \bar{l}' &= \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{k} \delta_j(z_i)(B_{F^*}(t(x_i) : \eta_j) - \log w_j), &(45)\\ &= \frac{1}{n} \sum_{i=1}^{n} \min_{j=1}^{k} (B_{F^*}(y_i : \eta_j) - \log w_j), &(46)\\ &= \mathrm{kmeans}_{F^*, \log w}(\mathcal{Y} : H), &(47) \end{aligned}$$

where $\mathcal{Y} = \{y_1 = t(x_1), ..., y_n = t(x_n(\}$ and $H = \{\eta_1, ..., \eta_k\}$.

Remark 1 *This is the argmin of Eq. 46 that gives the hidden component labels for the x_i's.*

Remark 2 *Observe that since $\forall i \in \{1, ..., k\}, -\log w_i \geq 0$ (since $w_i \leq 1$), we have the following additive dual Bregman divergence $B_{F^*}(y_i : \eta_j) - \log w_j > 0$ per cluster. Depending on the weights (e.g., $w \rightarrow 0$), we may have some empty clusters. In that case, the weight of a cluster is set to zero (and the component parameter is set to $\emptyset$ by convention). Note that it makes sense to consider ($\leq k$)-means instead of k-means in the sense that we would rather like to upper bound the maximum complexity of the model rather than precisely fixing it.*

Eq. 46 is precisely the loss function of a per-cluster additive Bregman k-means (see the appendix A) defined for the Legendre convex conjugate F^* of the log-normalizer F of the exponential family for the sufficient statistic points $\mathcal{Y} = \{y_i = t(x_i)\}_{i=1}^n$. It follows that *any* Bregman k-means heuristic decreases monotonically the loss function and reaches a local minimum (corresponding to a local maximum for the equivalent complete likelihood function). We can either use the batched Bregman Lloyd's k-means [9], the Bregman Hartigan and Wong's greedy cluster swap heuristic [23, 52], or the Kanungo et al. [25] $(9+\epsilon)$-approximation global swap approximation algorithm.

Remark 3 *The likelihood function L is equal to $e^{n\bar{l}}$. The average likelihood function $\bar{L}$ is defined by taking the geometric mean $\bar{L} = L^{\frac{1}{n}}$.*

The following section shows how to update the weights once the local convergence of the assignment-η of the k-MLE loop has been reached.

4 General k-MLE including mixture weight updates

When k-MLE with prescribed weights reaches a local minimum (see Eq. 44 and Eq. 46 and the appendix A), the current loss function is equal to

$$\bar{l} = \underbrace{\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k}\delta_j(z_i)(B_{F^*}(t(x_i) : \eta_j) - \log w_j)}_{\text{Minimized by additive Bregman } k\text{-means, see Appendix}} - \left(\frac{1}{n}\sum_{i=1}^{n} F^*(t(x_i)) + k(x_i)\right) \quad (48)$$

$$\bar{l} = \sum_{j=1}^{k} \alpha_j J_{F^*}(\mathcal{C}_j) - \alpha_j \log w_j - \left(\frac{1}{n}\sum_{i=1}^{n} F^*(t(x_i)) + k(x_i)\right) \quad (49)$$

where $\alpha_i = \frac{|\mathcal{C}_i|}{n}$ denotes the proportion of points assigned to the i-th cluster $\mathcal{C}_i$, and $\alpha_i J_{F^*}(\mathcal{C}_i)$ is the weighted Jensen diversity divergence of the cluster. In order to further minimize the average complete likelihood of Eq. 49, we update the mixture weights w_i's by minimizing the criterion:

$$\min_{w \in \Delta_k} \sum_{j=1}^{k} -\alpha_j \log w_j \quad (50)$$

$$= \min_{w \in \Delta_k} H^{\times}(\alpha : w), \quad (51)$$

where $H^\times(p : q) = -\sum_{i=1}^k p_i \log q_i$ denotes the Shannon cross-entropy, and Δ_k the $(k-1)$-dimensional probability simplex. The cross-entropy $H^\times(p : q)$ is minimized for $p = q$, and yields $H^\times(p, p) = H(p) = -\sum_{i=1}^k p_i \log p_i$, the Shannon entropy. Thus we update the weights by taking the relative proportion of points falling into the clusters:

$$\forall i \in \{1, ..., k\}, w_i \leftarrow \alpha_i. \tag{52}$$

After updated the weights, the average complete log-likelihood is

$$\bar{l} = \sum_{i=1}^k w_i J_{F^*}(\mathcal{C}_i) + H(w) - \left(\frac{1}{n} \sum_{i=1}^n F^*(t(x_i)) + k(x_i) \right). \tag{53}$$

We summarize the k-MLE algorithm in the boxed Algorithm 1.

Algorithm 1 Generic k-MLE for learning an exponential family mixture model.

Input:

$\mathcal{X}$	:	a set of n identically and independently distributed observations: $\mathcal{X} = \{x_1, ..., x_n\}$
F	:	log-normalizer of the exponential family, characterizing E_F
∇F	:	gradient of F for moment η-parameterization: $\eta = \nabla F(\theta)$
∇F^{-1}	:	functional inverse of the gradient of F for θ-parameterization: $\theta = \nabla F^{-1}(\eta)$
$t(x)$	:	the sufficient statistic of the exponential family
k	:	number of clusters

- 0. **Initialization**: $\forall i \in \{1, ..., k\}$, let $w_i = \frac{1}{k}$ and $\eta_i = t(x_i)$
  (Proper initialization is further discussed later on).

- 1. **Assignment**: $\forall i \in \{1, ..., n\}, z_i = \mathrm{argmin}_{j=1}^k B_{F^*}(t(x_i) : \eta_j) - \log w_j$.
  Let $\forall i \in \{1, ..., k\}$ $\mathcal{C}_i = \{x_j | z_j = i\}$ be the cluster partition: $\mathcal{X} = \cup_{i=1}^k \mathcal{C}_i$.
  (some clusters may become empty depending on the weight distribution)

- 2. **Update the η-parameters**: $\forall i \in \{1, ..., k\}, \eta_i = \frac{1}{|\mathcal{C}_i|} \sum_{x \in \mathcal{C}_i} t(x)$.
  (By convention, $\eta_i = \emptyset$ if $|\mathcal{C}_i| = 0$) **Goto step 1** unless local convergence of the complete likelihood is reached.

- 3. **Update the mixture weights**: $\forall i \in \{1, ..., k\}, w_i = \frac{1}{n} |\mathcal{C}_i|$.
  Goto step 1 unless local convergence of the complete likelihood is reached.

Output: An exponential family mixture model $m(x)$ (EFMM) parameterized in the natural coordinate system: $\forall i \in \{1, ..., k\}, \theta_i = (\nabla F)^{-1}(\eta_i) = \nabla F^*(\eta_i)$:

$$m(x) = \sum_{i=1}^k w_i p_F(x | \theta_i)$$

Remark 4 *Note that we can also do after the assignment step of data to clusters both (i) the mixture η-parameter update and (ii) the mixture w-weight update consecutively in a single iteration* *of the k-MLE loop. This corresponds to the Bregman hard expectation-maximization (Bregman Hard EM) algorithm described in boxed Algorithm 2. This Hard EM algorithm is straightforwardly implemented in legacy source codes by hardening the weight membership in the E-step of the EM. Hard EM was shown computationally efficient when learning mixtures of von-Mises Fisher (vMF) distributions [8]. Indeed, the log-normalizer F (used when computing densities) of vMF distributions requires to compute a modified Bessel function of the first kind [49], that is only invertible approximately using numerical schemes.*

Algorithm 2 Hard EM for learning an exponential family mixture model.

- 0. **Initialization**: $\forall i \in \{1, ..., k\}$, let $w_i = \frac{1}{k}$ and $\eta_i = t(x_i)$
  (Proper initialization is further discussed later on).
- 1. **Assignment**: $\forall i \in \{1, ..., n\}, z_i = \mathrm{argmin}_{j=1}^{k} B_{F^*}(t(x_i) : \eta_j) - \log w_j$.
  Let $\forall i \in \{1, ..., k\}$ $\mathcal{C}_i = \{x_j | z_j = i\}$ be the cluster partition: $\mathcal{X} = \cup_{i=1}^{k} \mathcal{C}_i$.
- 2. **Update the η-parameters**: $\forall i \in \{1, ..., k\}, \eta_i = \frac{1}{|\mathcal{C}_i|} \sum_{x \in \mathcal{C}_i} t(x)$.
- 3. **Update the mixture weights**: $\forall i \in \{1, ..., k\}, w_i = \frac{|\mathcal{C}_i|}{n}$.
- **Goto step 1** unless local convergence of the complete likelihood is reached.

We can also sparsify EM by truncating to the first D entries on each row (thus, we obtain a well-defined centroid per cluster for non-degenerate input). This is related to the sparse EM proposed in [35]. Degeneraties of the EM GMM is identified and discussed in [6]. Asymptotic convergence rate of the EM GMM is analyzed in [32].

There are many ways to initialize k-means [42]. Initialization shall be discussed in Section 6.

5 Speeding up k-MLE and Hard EM using Bregman NN queries

The proximity cells $\{\mathcal{V}_1, ..., \mathcal{V}_k\}$ induced by the cluster centers $\mathcal{C} = \{c_1, ..., c_k\}$ (in the η-coordinate system) are defined by:

$$\mathcal{V}_j = \{x \in \mathbb{X} \mid B_{F^*}(t(x) : \eta_j) - \log w_j \leq B_{F^*}(t(x) : \eta_l) - \log w_l, \forall l \in \{1, ..., k\} \backslash \{j\}\} \tag{54}$$

partitions the support $\mathbb{X}$ into a Voronoi diagram. It is precisely equivalent to the intersection of a Bregman Voronoi diagram for the dual log-normalizer F^* with additive weights [12] on the expectation parameter space $\mathbb{M} = \{\eta = \nabla F(\theta) \mid \theta \in \mathbb{N}\}$ with the hypersurface[4] $\mathbb{T} = \{t(x) \mid x \in \mathbb{X}\}$. For the case of Gaussian mixtures, the log-density of the joint distribution $w_i p_F(x; \mu_i, \Sigma_i)$ induces a partition of the space into an anisotropic weighted Voronoi diagram [27]. This is easily understood by taking *minus the log-density* of the Gaussian distribution (see Eq. 2):

$$-\log p(x; \mu_i, \Sigma_i) = \frac{1}{2} D_{\Sigma_i^{-1}}(x - \mu_i, x - \mu_i) + \frac{1}{2} \log |\Sigma_i| + \frac{d}{2} \log 2\pi, \tag{55}$$

[4] Note that there is only one global minimum for the distance $B_{F^*}(y : \eta)$ with $y \in \mathbb{T}$.

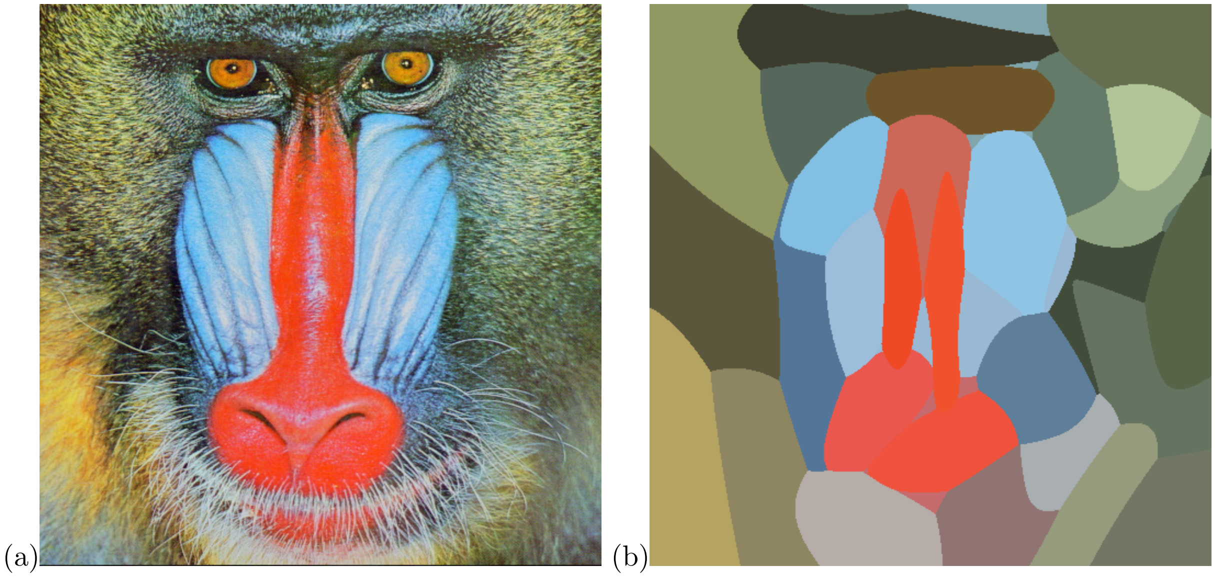


Figure 3: From the source color image (a), we buid a 5D GMM with $k = 32$ components, and color each pixel with the mean color of the anisotropic Voronoi cell it belongs to.

with M_Q the squared Mahalanobis distance $M_Q(x, y) = (x - y)^T Q(x - y)$. This is an additively weighted Bregman divergence with mass $m_i = \frac{1}{2}\log|\Sigma_i| + \frac{d}{2}\log 2\pi$ and generator $F_i(x) = \langle x, \Sigma_i^{-1} x\rangle$, the precision matrix (see the Appendix). Figure 3 displays the anisotropic Voronoi diagram [27] of a 5D xyRGB GMM restricted to the xy plane. We color each pixel with the mean color of the anisotropic Voronoi cell it belongs to.

When the order of the exponential family (i.e., number of parameters) is small (say, $D \leq 3$), we can compute explicitly this additively weighted Bregman Voronoi diagrams in the moment parameter space $\mathbb{M}$, and use proximity location data-structures designed for geometric partitions bounded by planar walls. Otherwise, we speed up the assignment step of k-MLE/Hard EM by using proximity location data-structures such as Bregman ball trees [45] or Bregman vantage point trees [40]. See also [1].

Besides Lloyd's batched k-means heuristic [31, 33, 19], we can also implement other k-means heuristic like the greedy Hartigan and Wong's swap [23, 52] in k-MLE that selects a point and optimally reassign it, or Kanungo et al. [25] global swap optimization, etc.

Remark 5 *The MLE equation $\hat{\eta} = \nabla F(\hat{\theta}) = \frac{1}{n}\sum_{i=1}^{n} t(x_i)$ may yield a transcendental equation. That is, when $(\nabla F)^{-1}$ is not available analytically (e.g., von Mises-Fisher family [8]), the convex conjutate F^* needs to be approximated by computing numerically the reciprocal gradient ∇F^{-1} (see Eq. 11). Sra [49] focuses on solving efficiently the MLE equation*[5] *for the von Mises-Fisher distributions.*

[5] See also, software R package `movMF`

6 Initializing k-MLE using k-MLE++

To complete the description of k-MLE of boxed Algorithm 1, it remains the problem to properly initializing k-MLE (step 0). One way to perform this initialization is to compute the global MLE parameter for the full set $\mathcal{X}$:

$$\hat{\eta} = \nabla F^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} t(x_i)\right), \tag{56}$$

and then consider the *restricted exponential family* of order $d \leq D$ with *restricted sufficient statistic* the first d components of full family statistic $(t_1(x), ..., t_d(x))$. We initialize the i-th cluster with $\eta_i^{(0)} = (t_1(x_i), ..., t_d(x_i), \hat{\eta}_{d+1}, ..., \hat{\eta}_D)$. For the case of multivariate Gaussians with $D = \frac{d(d+3)}{2}$, this amounts to compute the covariance matrix $\hat{\Sigma}$ of the full set and then set the translation parameter to x_i: $\eta_i^{(0)} = (x_i, -\frac{1}{2}(\hat{\Sigma} + x_i x_i^T))$ (see appendix B). This initialization is a heuristic with *no guaranteed performance* on the initial average complete log-likelihood $\bar{l}$ compared to the best one $\bar{l}^*$. Note that when $D = d$ (e.g., Poisson, Weibull, Rayleigh, isotropic Gaussian, etc.), we need to have distinct initializations so that instead of taking the global MLE, we rather split the data set into k groups of size $\frac{n}{k}$, and take the MLE of each group for initialization. A good geometric split is given by using a Voronoi partition diagram as follows: We run Bregman k-means on $\mathcal{Y}$ for the dual convex conjugate F^* and set the mixture parameters as the MLEs of clusters and the weights as the relative proportion of data in clusters. This corroborates an experimental observation by Banerjee et al. [9] that observes that clustering works experimentally best if we choose the dual Bregman divergence associated with the exponential family mixture sample set.

Let us further use the dual Bregman k-means interpretation of EM to perform this initialization efficiently. Assume uniform weighting of the mixtures. That is, $\forall i \in \{1, ..., k\}, w_i = \frac{1}{k}$.

Maximizing the average complete log-likelihood amounts to minimize (see Eq. 46):

$$\bar{l}'' = \frac{1}{n}\sum_{i=1}^{k} \min_{j=1}^{k} B_{F^*}(y_i = t(x_i) : \eta_j). \tag{57}$$

The likelihood function $L(x_1, ..., x_n|\theta, w)$ is

$$L = e^{-n\mathrm{kmeans}_{F^*}(\mathcal{C}) + n\log k + \sum_{i=1}^{n}(F^*(x_i)+k(x_i))}. \tag{58}$$

Thus for uniform mixture weights, the ratio between two different k-means optimization with respective cluster centers $\mathcal{C}$ and $\mathcal{C}'$ is:

$$\frac{L}{L'} = e^{-n(\mathrm{kmeans}_{F^*}(\mathcal{C}) - \mathrm{kmeans}_{F^*}(\mathcal{C}'))} \tag{59}$$

We can use the standard Bregman k-means++ initialization [2] on the convex conjugate F^* that gives probabilistically a guaranteed $O(\mu^{-2}\log k)$ performance, where μ is a constant factor to be explained below. The Bregman k-means++ algorithm is recalled in boxed Algorithm 3.

Let kmeans_F^* denote the optimal Bregman k-means average loss function for generator F. Bregman k-means++ [2] described in Algorithm 3 ensures that

$$\mathrm{kmeans}_F{}^*(\mathcal{Y} : \mathcal{C}) \leq \mathrm{kmeans}_F(\mathcal{Y} : \mathcal{C}) \leq \frac{8}{\mu^2}(2 + \log k)\mathrm{kmeans}_F{}^*(\mathcal{Y} : \mathcal{C}) \tag{60}$$

Algorithm 3 Bregman k-means++: probabilistically guarantees a good initialization.

```
• Choose first seed $\mathcal{C} = \{y_l\}$, for $l$ uniformly random in $\{1, ..., n\}$.
• For $i \leftarrow 2$ to $k$
    – Choose $c_i \in \{y_1, ..., y_n\}$ with probability
          $p_i = \frac{B_F(c_i : \mathcal{C})}{\sum_{i=1}^n B_F(y_i : \mathcal{C})} = \frac{B_F(\mathcal{Y} : \mathcal{C})}{\mathrm{kmeans}_F(\mathcal{Y} : \mathcal{C})}$,
      where $B_F(c : \mathcal{C}) = \min_{p \in \mathcal{C}} B_F(c : p)$.
    – Add selected seed to the initialization seed set: $\mathcal{C} \leftarrow \mathcal{C} \cup \{c_i\}$, and reiterate until $|\mathcal{C}| = k$.
```

The factor μ in the upper bound is related to the notion of μ-similarity that we now concisely explain. Observe that the squared Mahalanobis distance $M_Q(p, q) = (p - q)^T Q(p - q)$ satisfies the double triangle inequality:

$$M_Q(p, q) \leq 2(M_Q(p, r) + M_Q(r, q)). \tag{61}$$

A Bregman divergence is said to have the μ-similarity on a domain $\mathcal{Y}$ if there exists a positive definite matrix $Q \succ 0$ on $\mathcal{Y} = \mathrm{conv}(y_1, ..., y_n)$ and a real $0 < \mu \leq 1$ such that

$$\mu M_Q(p, q) \leq B_F(p : q) \leq M_Q(p, q) \tag{62}$$

Since a Bregman divergence can also be interpreted as the remainder of a Taylor expansion using the Lagrange error term:

$$B_F(p : q) = (p - q)^T \frac{\nabla^2 F(\epsilon_{pq})}{2} (p - q), \tag{63}$$

with ϵ_{pq} being a point on the line segment $[pq]$. It follows that by considering the Hessian $\nabla^2 F$ on a compact subset $\mathcal{Y} = \mathrm{conv}(y_1, ..., y_n)$, we get a bound [41] for μ as follows:

$$\mu = \min_{p,q \in \mathcal{Y}} \frac{\min_{y \in \mathcal{Y}} (p - q)^T \nabla^2 F(y)(p - q)}{\max_{y \in \mathcal{Y}} (p - q)^T \nabla^2 F(y)(p - q)}. \tag{64}$$

By considering a hyperrectangle bounding the convex hull $\mathcal{Y} = \mathrm{conv}(y_1, ..., y_n)$, it is usually easy to compute bounds for μ. See [2] for some examples.

The notion of μ-similarity also allows one to design fast proximity queries [1] based on the following two properties:

Approximately symmetric.

$$B_F(p : q) \leq \frac{1}{\mu} B_F(q, p) \tag{65}$$

Deficient triangle inequality.

$$B_F(p : q) \leq \frac{2}{\mu} (B_F(p : r) + B_F(q : r)) \tag{66}$$

For mixtures with prescribed but different non-zero weighting, we can bound the likelihood ratio using $w^+ = \max_i w_i \geq \frac{1}{k}$ and $w^- = \min_i w_i$. When mixture weights are unknown, we can further discretize weights by increments of size δ ($O(1/\delta^k)$ such weight combinations, where each combination gives rise to a fixed weighting) and choose the initialization that yields the best likelihood.

7 Concluding remarks and discussion

Banerjee et al. [9] proved that EM for learning exponential family mixtures amount to perform a dual Bregman soft clustering. Based on the duality between exponential families and Bregman divergences, we proposed k-MLE, a Bregman hard clustering in disguise. While k-MLE decreases monotonically the complete likelihood until it converges to a local minimum after a finite number of steps, EM monotonically decreases the expected complete likelihood and requires necessarily a prescribed stopping criterion. Because k-MLE uses hard membership of observations, it fits the doubly stochastic process of sampling mixtures (for which soft EM brings mathematical convenience).

Both k-MLE and EM are local search algorithm that requires to properly initialize the mixture parameters. We described k-MLE++, a simple initialization procedure that builds on Bregman k-means++ [2] to probabilistically guarantee an initialization not too far from the global optimum (in case of known weights). While we use Lloyd k-means [31] heuristic for minimizing the k-means loss, we can also choose other k-means heuristic to design a corresponding k-MLE. One possible choice is Hartigan's greedy swap [52] that can further improve the loss function when Lloyd's k-means is trapped into a local minimum. A local search technique such as Kanungo et al. swap [25] also guarantees a global $(9+\epsilon)$-approximation.

The MLE may yield degenerate situations when, say, one observation point is assigned to one component with weight close to one. For example, the MLE of one point for the normal distribution is degenerate as $\sigma \rightarrow 0$ (and $w \rightarrow 1$)), and the likelihood function tends to infinity. That is the unboundedness drawback of the MLE. See [48, 11] for further discussions on this topic including a penalization of the MLE to ensure boundedness.

Statistical mixtures with k components are generative models of overall complexity $k-1+kD$, where D is the order of the exponential family. An interesting future direction would be to compare mixture models versus a *single* multi-modal exponential family [16] (with implicit log-normalizer F). We did not address the model selection problem that consists in determining the appropriate number of components, nor the type of distribution family. Although there exists many criteria like the Akaike Information Criterion (AIC), model selection is a difficult problem since some distributions exhibit the indivisibility property that makes the selection process unstable. For example, a normal distribution can be interpreted as a sum of normal distributions: $\forall k \in \mathbb{N},\ N(\mu,\sigma^2) = \sum_{i=1}^k N\left(\frac{\mu}{k}, \frac{\sigma^2}{k}\right)$. From the practical point of view, it is better to overestimate k, and then perform mixture simplification using entropic clustering [20]. Belkin and Sinha [10] studied the polynomial complexity of learning a Gaussian mixture model.

We conclude by mentioning that it is still an active research topic to find good GMM learning algorithms in practice (e.g., see the recent entropy-based algorithm [43]).

Acknowledgments

FN (5793b870) would like to thank Joris Geessels for an early prototype in Python, Professor Richard Nock for stimulating discussions, and Professor Mario Tokoro and Professor Hiroaki Kitano for encouragements.

A k-Means with per-cluster additively weighted Bregman divergence

k-Means clustering asks to minimize the cost function kmeans($\mathcal{X} : \mathcal{C}$) by partitioning input set $\mathcal{X} = \{x_1, ..., x_n\}$ into k clusters using centers $\mathcal{C} = \{c_1, ..., c_k\}$, where

$$\text{kmeans}(\mathcal{X} : \mathcal{C}) = \frac{1}{n} \sum_{i=1}^{n} \min_{j=1}^{k} \|x_i - c_j\|^2. \tag{67}$$

There are several popular heuristics to minimize Eq. 67 like Lloyd's batched method [30] or Hartigan and Wong's swap technique [23]. Those iterative heuristics guarantee to decrease monotonically the k-means loss but can be trapped into a local minimum. In fact, solving for the global minimum kmeans$^*(\mathcal{X} : \mathcal{C})$ is NP-hard for general k (even on the plane) and for $k = 2$ and arbitrary dimension of datasets. Kanungo et al. [25] swap optimization technique guarantees a $(9 + \epsilon)$-approximation factor, for any $\epsilon > 0$.

Let us consider an additively weighted Bregman divergence B_{F_i,m_i} per cluster as follows:

$$B_{F_i,m_i}(p : q) = B_{F_i}(p : q) + m_i, \tag{68}$$

with m_i denoting the additive mass attached to a cluster center[6], and B_{F_i} the Bregman divergence induced by the Bregman generator F_i defined by

$$B_{F_i}(p : q) = F_i(p) - F_i(q) - \langle p - q, \nabla F_i(q) \rangle, \tag{69}$$

Remark 6 *For k-MLE, we consider all component distributions of the same exponential family E_F, and therefore all $F_i = F^*$'s are identical. We could have also considered different exponential families for the components but this would have burdened the paper with additional notations although it is of practical interest. For example, for the case of the multivariate Gaussian family, we can split the vector parameter part from the matrix parameter part, and write $F(\theta_{vi}, \theta_{Mi}) = F_{\theta_{Mi}}(\theta_{vi}) = F_i(\theta_{vi})$.*

Let us extend the Bregman batched Lloyd's k-means clustering [9] by considering the generalized k-means clustering loss function for a data set $\mathcal{Y} = \{y_1, ..., y_n\}$ and a set $\mathcal{C}$ of k cluster centers $\mathcal{C} = \{c_1, ..., c_k\}$:

$$\text{kmeans}(\mathcal{Y}, \mathcal{C}) = \min_{c_1,...,c_k} \frac{1}{n} \sum_{i=1}^{n} \min_{j=1}^{k} D_i(y_i : c_j). \tag{70}$$

Let us prove that the center-based Lloyd's k-means clustering algorithm monotonically decreases this loss function, and terminates after a finite number of iterations into a local optimum.

[6] In this paper, we have $m_i \geq 0$ by choosing $m_i = -\log w_i$ for $w_i < 1$, but this is not required.

- When $k=1$, the minimizer of $\mathrm{kmeans}(\mathcal{Y}, \mathcal{C}=\{c_1\})$ is the center of mass (always independent of the Bregman generator):

  $$c_1 = \frac{1}{n}\sum_{i=1}^n y_i = \bar{y}, \tag{71}$$

  and the Bregman information [9] is defined as the minimal 1-means loss function:

  $$\begin{aligned} \mathrm{kmeans}_{F_1,m_1}(\mathcal{Y},\{c_1\}) &= I_{F_1}(\mathcal{Y}) && (72)\\ &= \frac{1}{n}\sum_{i=1}^n F_1(y_i) - F_1(\bar{y}) + m_1, && (73)\\ &= m_1 + J_{F_1}(\mathcal{Y}), && (74)\end{aligned}$$

  where $\bar{y} = \frac{1}{n}\sum_{i=1}^n y_i$ and

  $$J_{F_1}(y_1,...,y_n) = \frac{1}{n}\sum_{i=1}^n F_1(y_i) - F_1(\bar{y}) \geq 0, \tag{75}$$

  denotes the Jensen diversity index [38].

- When $k \geq 2$, let $c_i^{(t)}$ denote the cluster center of the i-th cluster $\mathcal{C}_i^{(t)} \subset \mathcal{Y}$ of the partition $\mathcal{Y} = \cup_{i=1}^k \mathcal{C}_i^{(t)}$ at the t^{th} iteration. The generalized additively weighted Bregman k-means loss function can be rewritten as

  $$\mathrm{kmeans}_{F,m}(\mathcal{C}_1^{(t)},...,\mathcal{C}_k^{(t)} : c_1^{(t)},...,c_k^{(t)}) = \frac{1}{n}\sum_{i=1}^k \sum_{y\in\mathcal{C}_i^{(t)}} B_{F_i,m_i}(y:c_i). \tag{76}$$

  Since the assignment step allocates y_i to their closest cluster center $\mathrm{argmin}_{j=1}^k B_{F_i,m_i}(y_i : c_j)$, we have

  $$\mathrm{kmeans}_{F,m}(\mathcal{C}_1^{(t+1)},...,\mathcal{C}_k^{(t+1)} : c_1^{(t)},...,c_k^{(t)}) \leq \mathrm{kmeans}_{F,m}(\mathcal{C}_1^{(t)},...,\mathcal{C}_k^{(t)} : c_1^{(t)},...,c_k^{(t)}). \tag{77}$$

  Since the center relocation minimizes the average additively weighted divergence, we have

  $$\mathrm{kmeans}_{F,m}(\mathcal{C}_1^{(t+1)},...,\mathcal{C}_k^{(t+1)} : c_1^{(t+1)},...,c_k^{(t+1)}) \leq \mathrm{kmeans}_{F,m}(\mathcal{C}_1^{(t+1)},...,\mathcal{C}_k^{(t+1)}; c_1^{(t)},...,c_k^{(t)}). \tag{78}$$

  By iterating the assignment-relocation steps of k-means, and cascading the inequalities by transitivity, we get

  $$\mathrm{kmeans}_{F,m}(\mathcal{C}_1^{(t+1)},...,\mathcal{C}_k^{(t+1)} : c_1^{(t+1)},...,c_k^{(t+1)}) \leq \mathrm{kmeans}_{F,m}(\mathcal{C}_1^{(t)},...,\mathcal{C}_k^{(t)} : c_1^{(t)},...,c_k^{(t)}) \tag{79}$$

  Since the loss function is trivially lower bounded by $\frac{1}{n}\min_{i=1}^k m_i$ (and therefore always positive when all $m_i \geq 0$), we conclude that the generalized Bregman k-means converge to a local optimum, after a finite number[7] of iterations.

[7] We cannot repeat twice a partition.

Furthermore, the loss function can be expressed as

$$\begin{aligned}\mathrm{kmeans}_{F,m}(\mathcal{C}_1,...,\mathcal{C}_k : c_1,...,c_k) &= \frac{1}{n}\sum_{i=1}^{k}\sum_{y\in\mathcal{C}_i} B_{F_i,m_i}(y:c_i), && (80)\\ &= \sum_{i=1}^{k} w_i J_{F_i}(\mathcal{C}_i) + \sum_{i=1}^{k} w_i m_i, && (81)\end{aligned}$$

with $J_{F_i}(\mathcal{C}_i) = \frac{1}{|\mathcal{C}_i|}\sum_{y\in\mathcal{C}_i}^{n} F_i(y) - F_i(c_i) \geq 0$ (and $c_i = \frac{\sum_{y\in\mathcal{C}_i} y}{|\mathcal{C}_i|}$), and $w_i = \frac{|\mathcal{C}_i|}{n}$ for all $i \in \{1,...,k\}$, the cluster relative weights.

When all F_i are identical to some generator F, we have the following loss function:

$$\mathrm{kmeans}_{F,m} = \sum_{i=1}^{k} w_i J_F(\mathcal{C}_i) + \sum_{i=1}^{k} w_i m_i \qquad (82)$$

The celebrated k-means of Lloyd [30] minimizes the weighted within-cluster variances (for the Bregman quadratic generator $F(x) = \langle x, x\rangle$ inducing the squared Euclidean distance error) as shown in Eq. 81, with Bregman information:

$$\begin{aligned}J_F(\mathcal{Y}) &= \sum_{y\in\mathcal{Y}} \frac{1}{|\mathcal{Y}|}\|y-\bar{y}\|^2, && (83)\\ &= \sum_{y\in\mathcal{Y}} \frac{1}{|\mathcal{Y}|}\langle y-\bar{y}, y-\bar{y}\rangle, && (84)\\ &= \sum_{y\in\mathcal{Y}} \frac{1}{|\mathcal{Y}|}(\langle y,y\rangle - 2\langle\bar{y},y\rangle - \langle\bar{y},\bar{y}\rangle), && (85)\\ &= \sum_{y\in\mathcal{Y}} \frac{1}{|\mathcal{Y}|}\langle y,y\rangle - 2\left\langle \bar{y}, \underbrace{\sum_{y\in\mathcal{Y}} \frac{1}{|\mathcal{Y}|} y}_{\bar{y}} \right\rangle - \langle\bar{y},\bar{y}\rangle, && (86)\\ &= \frac{1}{|\mathcal{Y}|}\sum_{y\in\mathcal{Y}}\langle y,y\rangle - \langle\bar{y},\bar{y}\rangle = J_F(\mathcal{Y}), && (87)\end{aligned}$$

the variance. When all cluster generators are identical and have no mass, it is shown by Banerjee et al. [9] that the loss function can be equivalently rewritten as:

$$\begin{aligned}\mathrm{kmeans}_F(\mathcal{P}:\mathcal{C}) &= J_F(\mathcal{P}) - J_F(\mathcal{C}) = \sum_{i=1}^{k} w_i J_F(\mathcal{C}_i), && (88)\\ &= I_F(\mathcal{P}) - I_F(\mathcal{C}) && (89)\end{aligned}$$

Remark 7 *Note that we always have $\bar{c} = \bar{y}$. That is, the centroid $\bar{y}$ of set $\mathcal{Y}$ is equal to the barycenter $\bar{c}$ of the cluster centers $\mathcal{C}$ (with weights taken as the relative proportion of points falling within the clusters.*

Remark 8 *A multiplicatively weighted Bregman divergence $m_i B_{F_i}$ is mathematically equivalent to a Bregman divergence $B_{m_i F_i}$ for generator $m_i F_i$, provided that $m_i > 0$.*

As underlined in this proof, Lloyd's k-means [30] assignment-center relocation loop is a generic algorithm that extends to arbitrary divergences D_i guaranteeing unique average divergence minimizers, and the assignment/relocation process ensures that the associated k-means loss function decreases monotonically. Teboulle studied [51] generic center-based clustering optimization methods. It is however difficult to reach the global minimum since k-means is NP-hard, even when data set $\mathcal{Y}$ lies on the plane [53] for arbitrary k. In the worst case, k-means may take an exponential number of iterations to converge [53], even on the plane.

B Dual parameterization of the multivariate Gaussian (MVN) family

Let us explicit the dual θ-natural and η-moment parameterizations of the family of multivariate Gaussians. Consider the multivariate Gaussian probability density parameterized by a mean vector $\lambda_v = \mu$ and a covariance matrix $\lambda_M = \Sigma$.

$$p(x;\lambda) = \frac{1}{(2\pi)^{\frac{d}{2}}\sqrt{|\lambda_M|}} e^{-\frac{1}{2}(x-\lambda_v)^T \lambda_M^{-1}(x-\lambda_v)}, \tag{90}$$

$$= \exp\left(-\frac{1}{2}x^T\lambda_M^{-1}x + \lambda_v^T\lambda_M^{-1}x - \frac{1}{2}\lambda_v^T\lambda_M^{-1}\lambda_v - \frac{d}{2}\log 2\pi - \frac{1}{2}\log|\lambda_M|\right), \tag{91}$$

where the usual parameter is $\lambda = (\lambda_v, \lambda_M) = (\mu, \Sigma)$. Using the matrix cyclic trace property $-\frac{1}{2}x^T\lambda_M^{-1}x = \mathrm{tr}(-\frac{1}{2}xx^T\lambda_M^{-1})$ and the fact that $(\lambda_M^{-1})^T = \lambda_M^{-1}$, we rewrite the density as follows:

$$p(x;\lambda) = \exp\left(\langle x, \lambda_M^{-1}\lambda_v\rangle + \langle -\frac{1}{2}xx^T, \lambda_M^{-1}\rangle - \left(\frac{1}{2}\lambda_v^T\lambda_M^{-1}\lambda_v + \frac{d}{2}\log 2\pi + \frac{1}{2}\log|\lambda_M|\right)\right), \tag{92}$$

where the inner product of vector is $\langle v_1, v_2\rangle = v_1^T v_2$ and the inner product of matrices is $\langle M_1, M_2\rangle = \mathrm{tr}(M_1^T M_2)$. Thus we define the following canonical terms:

- sufficient statistics: $t(x) = (x, -\frac{1}{2}xx^T)$,
- auxiliary carrier measure: $k(x) = 0$,
- natural parameter: $\theta = (\theta_v, \theta_M) = (\lambda_M^{-1}\lambda_v, \lambda_M^{-1})$.
- log-normalizer expressed in the λ-coordinate system:

$$F(\lambda) = \frac{1}{2}\lambda_v^T\lambda_M^{-1}\lambda_v + \frac{d}{2}\log 2\pi + \frac{1}{2}\log|\lambda_M| \tag{93}$$

Since $\lambda_v = \theta_M^{-1}\theta_v$ (and $\lambda_v^T = \theta_v^T\theta_M^{-1}$) and $\log|\lambda_M| = -\log|\theta_M|$, we express the log-normalizer in the θ-coordinate system as follows:

$$F(\theta) = \frac{1}{2}\theta_v^T\theta_M^{-1}\theta_v - \frac{1}{2}\log|\theta_M| + \frac{d}{2}\log 2\pi \tag{94}$$

Since the derivative of the log determinant of a symmetric matrix is $\nabla_X \log |X| = X^{-1}$ and the derivative of an inverse matrix trace [44]:

$$\nabla_X \mathrm{tr}(AX^{-1}B) = -(X^{-1}BAX^{-1})^T \tag{95}$$

(applied to $\frac{1}{2}\mathrm{tr}(\theta_v^T \theta_M^{-1} \theta_v) = -\frac{1}{2}(\theta_M^{-1}\theta_v\theta_v^T\theta_M^{-1}))$, we calculate the gradient ∇F of the log-normalizer as

$$\nabla F(\theta) = (\nabla_{\theta_v} F(\theta), \nabla_{\theta_M} F(\theta)) \tag{96}$$

with

$$\begin{aligned}
\eta_v &= \nabla_{\theta_v} F(\theta) = \theta_M^{-1}\theta_v, && (97)\\
&= E[x] = \mu, && (98)\\
\eta_M &= \nabla_{\theta_M} F(\theta) = -\frac{1}{2}(\theta_M^{-1}\theta_v)(\theta_M^{-1}\theta_v)^T - \frac{1}{2}\theta_M^{-1}, && (99)\\
&= E\left[-\frac{1}{2}xx^T\right] = -\frac{1}{2}(\mu\mu^T + \Sigma), && (100)
\end{aligned}$$

where $\eta = \nabla F(\theta) = (\eta_v, \eta_M)$ denotes the dual moment parameterization of the Gaussian.

It follows that the Kullback-Leibler divergence of two multivariate Gaussians is

$$\begin{aligned}
\mathrm{KL}(p(x;\lambda_1) : p(x;\lambda_2)) &= B_F(\theta_2 : \theta_1), && (101)\\
&= \frac{1}{2}\left(\mathrm{tr}(\Sigma_2^{-1}\Sigma_1) - \log|\Sigma_1\Sigma_2^{-1}| + (\mu_2-\mu_1)^T\Sigma_2^{-1}(\mu_2-\mu_1)\right). && (102)
\end{aligned}$$

Note that the Kullback-Leibler divergence of multivariate Gaussian distributions [17] can be decomposed as the sum of a Burg matrix divergence (Eq. 106) with a squared Mahalanobis distance (Eq. 106) (both being Bregman divergences):

$$\begin{aligned}
\mathrm{KL}(p_F(x|\mu_1,\Sigma_1) : p_F(x|\mu_2,\Sigma_2) &= \frac{1}{2}\left(\mathrm{tr}(\Sigma_2^{-1}\Sigma_1) - \log|\Sigma_1\Sigma_2^{-1}| + (\mu_2-\mu_1)^T\Sigma_2^{-1}(\mu_2-\mu_1)\right) && (103)\\
&= \frac{1}{2}B(\Sigma_1,\Sigma_2) + \frac{1}{2}M_{\Sigma_2^{-1}}(\mu_1,\mu_2), && (104)
\end{aligned}$$

with

$$\begin{aligned}
B(\Sigma_1 : \Sigma_2) &= \mathrm{tr}(\Sigma_1\Sigma_2^{-1}) - \log|\Sigma_1\Sigma_2^{-1}| - d, && (105)\\
M_{\Sigma_2^{-1}}(\mu_1,\mu_2) &= (\mu_1-\mu_2)^T\Sigma_2^{-1}(\mu_1-\mu_2). && (106)
\end{aligned}$$

To compute the functional inverse of the gradient, we write:

$$\theta = \nabla F^{-1}(\eta) = \nabla F^*(\eta). \tag{107}$$

Since $\eta_M = -\frac{1}{2}(\eta_v\eta_v^T + \theta_M^{-1})$, we have:

$$\begin{aligned}\theta_M &= (-2\eta_M - \eta_v\eta_v^T)^{-1}, & (108)\\ \theta_v &= (-2\eta_M - \eta_v\eta_v^T)^{-1}\eta_v. & (109)\end{aligned}$$

Finally, we get the Legendre convex conjugate $F^*(\eta)$ as:

$$\begin{aligned}F^*(\eta) &= \langle \nabla F^*(\eta), \eta\rangle - F(\nabla F^*(\eta)), & (110)\\ &= -\frac{1}{2}\log(1 + 2\eta_v^T\eta_M^{-1}\eta_v) - \frac{1}{2}\log|-\eta_M| - \frac{d}{2}\log(\pi e). & (111)\end{aligned}$$

C k-MLE for Gaussian Mixture Models (GMMs)

We explicit k-MLE for Gaussian mixture models on the usual (μ, Σ) parameters in Algorithm 4.

The k-MLE++ initialization for the GMM is reported in Algorithm 5.

D Rayleigh Mixture Models (RMMs)

We instantiate the soft Bregman EM, hard EM, k-MLE, and k-MLE++ for the Rayleigh distributions, a sub-family of Weibull distributions.

A Rayleigh distribution has probability density $\frac{x}{\sigma^2}e^{-\frac{x^2}{2\sigma^2}}$ where $\sigma \in \mathbb{R}^+$ denotes the *mode* of the distribution, and $x \in \mathbf{X} = \mathbb{R}^+$ the support. The Rayleigh distributions form a 1-order univariate exponential family ($D = d = 1$). Re-writing the density in the canonical form $e^{-\frac{x^2}{2\sigma^2}+\log x - 2\log\sigma}$, we deduce that $t(x) = x^2$, $\theta = -\frac{1}{2\sigma^2}$, $k(x) = \log x$, and $F(\sigma^2) = \log\sigma^2 = \log -\frac{1}{2\theta} = -\log(-2\theta) = F(\theta)$. Thus $\nabla F(\theta) = -\frac{1}{\theta} = \eta$ and $F^*(\eta) = \langle\theta, \eta\rangle - F(\theta) = -1 + \log\frac{2}{\eta}$. The natural parameter space is $\mathbb{N} = \mathbb{R}^-$ and the moment parameter space is $\mathbb{M} = \mathbb{R}^+$ (with $\eta = 2\sigma^2$). We check that conjugate gradients are reciprocal of each other since $\nabla F^*(\eta) = -\frac{1}{\eta} = \theta$, and we have $\nabla^2 F(\theta)\nabla^2 G(\eta) = \frac{1}{\theta^2}\frac{1}{\eta^2} = 1$ (i.e, dually orthogonal coordinate system) with $\nabla^2 F(\theta) = \frac{1}{\theta^2}$ and $\nabla^2 F^*(\eta) = \frac{1}{\eta^2}$.

Rayleigh mixtures are often used in ultrasound imageries [47].

D.1 EM as a Soft Bregman clustering algorithm

Following Banerjee et al. [9], we instantiate the Bregman soft clustering for the convex conjugate $F^*(\eta) = -1 + \log\frac{2}{\eta}$, $t(x) = x^2$ and $\eta = 2\sigma^2$. The Rayleigh density expressed in the η-parameterization yields $p(x;\sigma) = p(x;\eta) = \frac{2x}{\eta}e^{-\frac{2x^2}{\eta}}$.

Expectation. Soft membership for all observations $x_1, ..., x_n$:

$$\forall 1 \leq i \leq n, 1 \leq j \leq k,\ w_{i,j} = \frac{w_j p(x_i;\theta_j)}{\sum_{l=1}^k w_l p(x_i;\theta_l)}, \quad (112)$$

(We can use any of the equivalent σ, θ or η parameterizations for calculating the densities.)

Algorithm 4 k-MLE for learning a GMM.

Input:

X	:	a set of n independent and identically distributed distinct observations: $X = \{x_1, ..., x_n\}$
k	:	number of clusters

- 0. **Initialization**:
  - Calculate global mean $\bar{\mu}$ and global covariance matrix $\bar{\Sigma}$:

    $$\begin{aligned}\bar{\mu} &= \frac{1}{n}\sum_{i=1}^{k} x_i,\\ \bar{\Sigma} &= \frac{1}{n}\sum_{i=1}^{k} x_i x_i^T - \bar{\mu}\bar{\mu}^T\end{aligned}$$

  - $\forall i \in \{1, ..., k\}$, initialize the ith seed as $(\mu_i = x_i, \Sigma_i = \bar{\Sigma})$.
- 1. **Assignment**:

  $$\forall i \in \{1, ..., n\}, z_i = \mathrm{argmin}_{j=1}^{k} M_{\Sigma_i^{-1}}(x - \mu_i, x - \mu_i) + \log|\Sigma_i| - 2\log w_i$$

  with $M_{\Sigma_i^{-1}}(x - \mu_i, x - \mu_i)$ the squared Mahalanobis distance: $M_Q(x, y) = (x - y)^T Q(x - y)$.

  Let $\mathcal{C}_i = \{x_j | z_j = i\}, \forall i \in \{1, ..., k\}$ be the cluster partition: $X = \cup_{i=1}^{k} \mathcal{C}_i$.
  (Anisotropic Voronoi diagram [27])
- 2. **Update the parameters**:

  $$\forall i \in \{1, ..., k\}, \mu_i = \frac{1}{|\mathcal{C}_i|}\sum_{x \in \mathcal{C}_i} x, \Sigma_i = \frac{1}{|\mathcal{C}_i|}\sum_{x \in \mathcal{C}_i} x x^T - \mu_i \mu_i^T$$

  Goto step 1 unless local convergence of the complete likelihood is reached.
- 3. **Update the mixture weights**: $\forall i \in \{1, ..., k\}, w_i = \frac{1}{n}|\mathcal{C}_i|$.
  Goto step 1 unless local convergence of the complete likelihood is reached.

Algorithm 5 k-MLE for GMM:

```
- Choose first seed $\mathcal{C} = \{y_l\}$, for $l$ uniformly random in $\{1, ..., n\}$.
- For $i \leftarrow 2$ to $k$
    - Choose $c_i = (\mu_i, \Sigma_i)$ with probability
          $\frac{B_{F*}(c_i : \mathcal{C})}{\sum_{i=1}^n B_{F*}(y_i : \mathcal{C})} = \frac{B_{F^*}(\mathcal{Y} : \mathcal{C})}{\mathrm{kmeans}_{F^*}(\mathcal{Y} : \mathcal{C})},$
      where $B_{F^*}(c : \mathcal{P}) = \min_{p \in \mathcal{P}} B_{F^*}(c : p)$.
          $F^*(\mu, \Sigma) = -\frac{1}{2} \log\left(1 - \mu^T(\mu\mu^T + \Sigma)^{-1}\mu\right) - \frac{1}{2}\log|\mu^T\mu + \Sigma| - \frac{d}{2}\log 2\pi - d$
    - Add selected seed to the initialization seed set: $\mathcal{C} \leftarrow \mathcal{C} \cup \{c_i\}$.
```

Maximization. Barycenter in the moment parameterization:

$$\begin{aligned} \forall 1 \leq j \leq k,\ \eta_j &= \frac{\sum_{i=1}^n w_{i,j} t(x_i)}{\sum_{l=1}^n w_{l,j}}, && (113) \\ \sigma_j &= \sqrt{\frac{1}{2} \frac{\sum_{i=1}^n w_{i,j} x_i^2}{\sum_{l=1}^n w_{l,j}}} && (114) \end{aligned}$$

D.2 k-Maximum Likelihood Estimators

The associated Bregman divergence for the convex conjugate generator of the Rayleigh distribution log-normalizer is

$$\begin{aligned} B_{F^*}(\eta_1 : \eta_2) &= F^*(\eta_1) - F^*(\eta_2) - \langle \eta_1 - \eta_2, \nabla F^*(\eta_2) \rangle, && (115) \\ &= -1 + \log \frac{2}{\eta_1} + 1 - \log \frac{2}{\eta_2} - (\eta_1 - \eta_2)(-1/\eta_2), && (116) \\ &= \frac{\eta_1}{\eta_2} + \log \frac{\eta_2}{\eta_1} - 1, && (117) \\ &= \mathrm{IS}(\eta_1 : \eta_2) && (118) \end{aligned}$$

This is the Itakura-Saito divergence IS (indeed, F^* is equivalent modulo affine terms to $-\log \eta$, the Burg entropy).

1. Hard assignment.

$$\forall 1 \leq i \leq n, z_i = \mathrm{argmin}_{1 \leq j \leq k} \mathrm{IS}(x_i^2 : \eta_j) - \log w_j$$

Voronoi partition into clusters:

$$\forall 1 \leq j \leq k, \mathcal{C}_j = \{x_i \mid \mathrm{IS}(x_i^2 : \eta_j) - \log w_j \leq \mathrm{IS}(x_i^2 : \eta_l) - \log w_l \forall l \neq j\}$$

2. η-parameter update.

$$\forall 1 \leq j \leq k, \eta_j \leftarrow \frac{1}{|\mathcal{C}_j|} \sum_{x \in \mathcal{C}_j} x^2$$

$$\forall 1 \leq j \leq k, \sigma_j = \sqrt{\frac{1}{2}\eta_j}$$

Go to 1. until (local) convergence is met.

weight update.

$$\forall 1 \leq j \leq k, w_j = \frac{|\mathcal{C}_j|}{n}$$

Go to 1. until (local) convergence is met.

Note that k-MLE does also model selection as it may decrease the number of clusters in order to improve the complete log-likelihood. If initialization is performed using random point and uniform weighting, the first iteration ensures that all Voronoi cells are non-empty.

D.3 k-MLE++

A good initialization for Rayleigh mixture models is done as follows: Compute the order statistics for the $\frac{n}{k}$, $\frac{2n}{k}$, $\frac{(k-1)n}{k}$-th elements (in overall $O(n \log k)$-time). Those pivot elements split the set $\mathcal{X}$ into k groups $\mathcal{X}_1, ..., \mathcal{X}_k$ of size $\frac{n}{k}$, on which we estimate the MLEs.

The k-MLE++ initialization is built from the Itakura-Saito divergence:

$$\mathrm{IS}(\eta_1 : \eta_2) = \frac{\eta_1}{\eta_2} + \log \frac{\eta_2}{\eta_1} - 1$$

k-MLE++:

- Choose first seed $\mathcal{C} = \{y_l\}$, for l uniformly random in $\{1, ..., n\}$.
- For $i \leftarrow 2$ to k
  - Choose $c_i \in y_1 = x_1^2, ..., y_n = x_n^2$ with probability

    $$\frac{\mathrm{IS}(c_i : \mathcal{C})}{\sum_{i=1}^n \mathrm{IS}(y_i : \mathcal{C})}$$

  - Add selected seed to the initialization seed set: $\mathcal{C} \leftarrow \mathcal{C} \cup \{c_i\}$.

E Notations

Exponential family:	
$\langle x, y\rangle$	inner product (e.g., $x^\top y$ for vectors, $\mathrm{tr}(Y^\top X)$ for matrices)
$p_F(x;\theta) = e^{\langle t(x),\theta\rangle - F(\theta) + k(x)}$	Exponential distribution parameterized using the θ-coordinate system
$\mathbb{X}$	support of the distribution family ($\{x \mid p_F(x;\theta) > 0\}$)
d	dimension of the support $\mathbb{X}$ (univariate versus multivariate)
D	dimension of the natural parameter space (uniparameter versus multiparameter)
$t(x)$	sufficient statistic ($\hat{\eta} = \frac{1}{n}\sum_{i=1}^n t(x_i)$)
$k(x)$	auxiliary carrier term
F	log-normalizer, log-Laplace, cumulant function ($F : \mathbb{N} \to \mathbb{R}$)
∇F	gradient of the log-normalizer (for moment η-parameterization)
$\nabla^2 F$	Hessian of the log-normalizer (Fisher information matrix, SPD: $\nabla^2 F(\theta) \succ 0$)
F^*	Legendre convex conjugate
Distribution parameterization:	
θ	canonical natural parameter
$\mathbb{N}$	natural parameter space
η	canonical moment parameter
$\mathbb{M}$	moment parameter space
λ	usual parameter
$\mathbb{L}$	usual parameter space
$p_F(x;\lambda)$	density or mass function using the usual λ-parameterization
$p_F(x;\eta)$	density or mass function using the usual moment parameterization
Mixture:	
m	mixture model
Δ_k	closed probability $(d-1)$-dimensional simplex
$H(w)$	Shannon entropy $-\sum_{i=1}^d w_i \log w_i$ (with $0\log 0 = 0$ by convention)
$H^\times(p:q)$	Shannon cross-entropy $-\sum_{i=1}^d p \log q$
w_i	mixture weights (positive such that $\sum_{i=1}^k w_i = 1$)
θ_i	mixture component natural parameters
η_i	mixture component moment parameters
$\tilde{m}$	estimated mixture
k	number of mixture components
Ω	mixture parameters
Clustering:	
$\mathcal{X} = \{x_1, ..., x_n\}$	sample (observation) set
$\lvert\mathcal{X}\rvert, \lvert\mathcal{C}\rvert$	cardinality of sets: n for the observations, k for the cluster centers
$z_1, ..., z_n$	Hidden component labels
$\mathcal{Y} = \{y_1 = t(x_1), ..., y_n = t(x_n)\}$	sample sufficient statistic set

$L(x_1,...,x_n;\theta)$	likelihood function
$\hat{\theta},\hat{\eta},\hat{\lambda}$	maximum likelihood estimates
$w_{i,j}$	soft weight for x_i in cluster/component $\mathcal{C}_j$ (w_j,θ_j)
i	index on the sample set $x_1,...,x_i,...,x_n$
j	index on the mixture parameter set $\theta_1,...,\theta_j,...,\theta_k$
$\mathcal{C}$	cluster partition
$c_1,...,c_k$	cluster centers
$\alpha_1,...,\alpha_k$	cluster proportion size
B_F	Bregman divergence with generator F:

$$\begin{aligned} B_F(\theta_2,\theta_1) &= \mathrm{KL}(p_F(x:\theta_1):p_F(x:\theta_2)) \\ &= B_{F^*}(\eta_1,\eta_2) \\ &= F(\theta_2)+F^*(\eta_1)-\langle\eta_1,\theta_2\rangle \end{aligned}$$

J_F	Jensen diversity index: $J_F(p_1,...,p_n;w_1,...,w_n)=\sum_{i=1}^n w_iF(p_i)-F(\sum_{i=1}^n w_ip_i)\geq 0$
Evaluation criteria:	
$\bar{l}_F$	average incomplete log-likelihood: $\bar{l}_F(x_1,...,x_n)=\frac{1}{n}\sum_{i=1}^n\log\sum_{j=1}^k w_jp_F(x_i;\theta_j)$
$\bar{l}'_F$	average complete log-likelihood $\bar{l}'_F(x_1,...,x_n)=\frac{1}{n}\sum_{i=1}^n\log w_{z_i}p_F(x_i;\theta_{z_i})$
$\bar{L}_F$	geometric average incomplete likelihood: $\bar{L}_F(x_1,...,x_n)=e^{\bar{l}_F(x_1,...,x_n)}$
$\bar{L}'_F$	geometric average complete likelihood: $\bar{L}'_F(x_1,...,x_n)=e^{\bar{l}'_F(x_1,...,x_n)}$
kmeans_F	average k-means loss function (average divergence to the closest center)

$$\begin{aligned} \mathrm{kmeans}_F(\mathcal{X},\mathcal{C}) &= \frac{1}{n}\sum_{i=1}^n B_F(x_i:\mathcal{C}) \\ &= \frac{1}{n}\sum_{j=1}^k\sum_{x\in\mathcal{C}_j}B_F(x:c_j) \\ &= \sum_{j=1}^k w_jJ_F(\mathcal{C}_j) \\ &= J_F(\mathcal{X})-J_F(\mathcal{C}) \end{aligned}$$

$\mathrm{kmeans}_{F,m}$	average k-means loss function with respect to additive Bregman divergences

References

[1] Amirali Abdullah, John Moeller, and Suresh Venkatasubramanian. Approximate Bregman near neighbors in sublinear time: Beyond the triangle inequality. *CoRR*, abs/1108.0835, 2011.

[2] Marcel R. Ackermann and Johannes Blömer. Bregman clustering for separable instances. In *Scandinavian Workshop on Algorithm Theory (SWAT)*, pages 212–223, 2010.

[3] Shun-ichi Amari and Hiroshi Nagaoka. *Methods of Information Geometry.* Oxford University Press, 2000.

[4] Tahir Amin, Mehmet Zeytinoglu, and Ling Guan. Application of Laplacian mixture model to image and video retrieval. *IEEE Transactions on Multimedia*, 9(7):1416–1429, 2007.

[5] Yali Amit and Alain Trouvé. Generative models for labeling multi-object configurations in images. In *Toward Category-Level Object Recognition*, pages 362–381, 2006.

[6] Cédric Archambeau, John Aldo Lee, and Michel Verleysen. On convergence problems of the EM algorithm for finite Gaussian mixtures. In *European Symposium on Artificial Neural Networks (ESANN)*, pages 99–106, 2003.

[7] Arindam Banerjee, Inderjit Dhillon, Joydeep Ghosh, and Srujana Merugu. An information theoretic analysis of maximum likelihood mixture estimation for exponential families. In *Proceedings of the twenty-first international conference on Machine learning*, ICML, pages 57–64, New York, NY, USA, 2004. ACM.

[8] Arindam Banerjee, Inderjit S. Dhillon, Joydeep Ghosh, and Suvrit Sra. Clustering on the unit hypersphere using von Mises-Fisher distributions. *Journal of Machine Learning Research*, 6:1345–1382, December 2005.

[9] Arindam Banerjee, Srujana Merugu, Inderjit S. Dhillon, and Joydeep Ghosh. Clustering with Bregman divergences. *Journal of Machine Learning Research*, 6:1705–1749, 2005.

[10] Mikhail Belkin and Kaushik Sinha. Polynomial learning of distribution families. In *Foundations of Computer Science (FOCS)*, pages 103–112, 2010.

[11] Christophe Biernacki. Degeneracy in the maximum likelihood estimation of univariate Gaussian mixtures for grouped data and behaviour of the EM algorithm. *Scandinavian Journal of Statistics*, 34(3):569–586, 2007.

[12] Jean-Daniel Boissonnat, Frank Nielsen, and Richard Nock. Bregman Voronoi diagrams. *Discrete and Computational Geometry*, 44(2):281–307, April 2010.

[13] Lawrence D. Brown. *Fundamentals of statistical exponential families: with applications in statistical decision theory.* Institute of Mathematical Statistics, Hayworth, CA, USA, 1986. available on-line from Project Euclid.

[14] Chad Carson, Serge Belongie, Hayit Greenspan, and Jitendra Malik. Blobworld: Image segmentation using Expectation-Maximization and its application to image querying. *IEEE Transactions Pattern Analysis and Machine Intelligence*, 24(8):1026–1038, 2002.

[15] Jean-François Castet and Joseph H. Saleh. Single versus mixture Weibull distributions for nonparametric satellite reliability. *Reliability Engineering & System Safety*, 95(3):295 – 300, 2010.

[16] Loren Cobb, Peter Koppstein, and Neng Hsin Chen. Estimation and moment recursion relations for multimodal distributions of the exponential family. *Journal of the American Statistical Association*, 78(381):124–130, 1983.

[17] Jason V. Davis and Inderjit S. Dhillon. Differential entropic clustering of multivariate Gaussians. In Bernhard Scholkopf, John Platt, and Thomas Hoffman, editors, *Neural Information Processing Systems (NIPS)*, pages 337–344. MIT Press, 2006.

[18] Arthur Pentland Dempster, Nan M. Laird, and Donald B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

[19] Edward W. Forgy. Cluster analysis of multivariate data: efficiency vs interpretability of classifications. *Biometrics*, 1965.

[20] Vincent Garcia and Frank Nielsen. Simplification and hierarchical representations of mixtures of exponential families. *Signal Processing (Elsevier)*, 90(12):3197–3212, 2010.

[21] Hayit Greenspan, Jacob Goldberger, and Arnaldo Mayer. Probabilistic space-time video modeling via piecewise GMM. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26:384–396, March 2004.

[22] Leonard R. Haff, Peter T. Kim, Ja-Yong Koo, and Donald St. P. Richards. Minimax estimation for mixtures of Wishart distributions. *Annals of Statistics*, 39(arXiv:1203.3342 (IMS-AOS-AOS951)):3417–3440, Mar 2012.

[23] John A. Hartigan and Manchek A. Wong. Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108, 1979.

[24] Kittipat Kampa, Erion Hasanbelliu, and Jose Principe. Closed-form Cauchy-Schwarz PDF divergence for mixture of Gaussians. In *Proceeding of the International Joint Conference on Neural Networks (IJCNN)*, pages 2578 – 2585, 2011.

[25] Tapas Kanungo, David M. Mount, Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman, and Angela Y. Wu. A local search approximation algorithm for k-means clustering. *Computational Geometry*, 28(2-3):89–112, 2004.

[26] Michael Kearns, Yishay Mansour, and Andrew Y. Ng. An information-theoretic analysis of hard and soft assignment methods for clustering. In *Proceedings of the Thirteenth conference on Uncertainty in artificial intelligence*, UAI, pages 282–293, 1997.

[27] François Labelle and Jonathan Richard Shewchuk. Anisotropic Voronoi diagrams and guaranteed-quality anisotropic mesh generation. In *Proceedings of the nineteenth annual symposium on Computational geometry*, SCG '03, pages 191–200, New York, NY, USA, 2003. ACM.

[28] Jia Li and Hongyuan Zha. Two-way Poisson mixture models for simultaneous document classification and word clustering. *Computational Statistics & Data Analysis*, 50(1):163–180, 2006.

[29] Meizhu Liu, Baba C. Vemuri, Shun-ichi Amari, and Frank Nielsen. Shape retrieval using hierarchical total Bregman soft clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2012.

[30] Stuart P. Lloyd. Least squares quantization in PCM. Technical report, Bell Laboratories, 1957.

[31] Stuart P. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, IT-28(2):129–137, March 1982.

[32] Jinwen Ma, Lei Xu, and Michael I. Jordan. Asymptotic convergence rate of the EM algorithm for Gaussian mixtures. *Neural Computation*, 12(12):2881–2907, 2001.

[33] James B. MacQueen. Some methods of classification and analysis of multivariate observations. In L. M. Le Cam and J. Neyman, editors, *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*. University of California Press, Berkeley, CA, USA, 1967.

[34] Geoffrey McLachlan and David Peel. *Finite Mixture Models*. Wiley Series in Probability and Statistics. Wiley-Interscience, 1 edition, October 2000.

[35] Radford M. Neal and Geoffrey E. Hinton. A view of the EM algorithm that justifies incremental, sparse, and other variants. In M. I. Jordan, editor, *Learning in Graphical Models*. Kluwer, 1998.

[36] Frank Nielsen and Vincent Garcia. Statistical exponential families: A digest with flash cards, 2009. arXiv.org:0911.4863.

[37] Frank Nielsen and Richard Nock. Clustering multivariate normal distributions. In Frank Nielsen, editor, *Emerging Trends in Visual Computing*, pages 164–174. Springer-Verlag, Berlin, Heidelberg, 2009.

[38] Frank Nielsen and Richard Nock. Sided and symmetrized Bregman centroids. *IEEE Transactions on Information Theory*, 55(6):2048–2059, June 2009.

[39] Frank Nielsen and Richard Nock. Entropies and cross-entropies of exponential families. In *International Conference on Image Processing (ICIP)*, pages 3621–3624, 2010.

[40] Frank Nielsen, Paolo Piro, and Michel Barlaud. Bregman vantage point trees for efficient nearest neighbor queries. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 878–881, New York City, USA, June 2009. IEEE.

[41] Richard Nock, Panu Luosto, and Jyrki Kivinen. Mixed Bregman clustering with approximation guarantees. In *Proceedings of the European conference on Machine Learning and Knowledge Discovery in Databases*, pages 154–169, Berlin, Heidelberg, 2008. Springer-Verlag.

[42] José M. Pena, José A. Lozano, and Pedro Larranaga. An empirical comparison of four initialization methods for the k-means algorithm. *Pattern Recognition Letters*, 20(10):1027–1040, October 1999.

[43] Antonio Peñalver and Francisco Escolano. Entropy-based incremental variational Bayes learning of Gaussian mixtures. *IEEE Transactions on Neural Network and Learning Systems*, 23(3):534–540, 2012.

[44] K. B. Petersen and M. S. Pedersen. The matrix cookbook, 2008-2012.

[45] Paolo Piro, Frank Nielsen, and Michel Barlaud. Tailored Bregman ball trees for effective nearest neighbors. In *European Workshop on Computational Geometry (EuroCG)*, LORIA, Nancy, France, March 2009. IEEE.

[46] Loïs Rigouste, Olivier Cappé, and François Yvon. Inference and evaluation of the multinomial mixture model for text clustering. *Information Processing and Management*, 43(5):1260–1280, January 2007.

[47] Jose Seabra, Francesco Ciompi, Oriol Pujol, Josepa Mauri, Petia Radeva, and Joao Sanchez. Rayleigh mixture model for plaque characterization in intravascular ultrasound. *IEEE Transaction on Biomedical Engineering*, 58(5):1314–1324, 2011.

[48] Hichem Snoussi and Ali Mohammad-Djafari. Penalized maximum likelihood for multivariate Gaussian mixture. *Aip Conference Proceedings*, pages 36–46, 2001.

[49] Suvrit Sra. A short note on parameter approximation for von Mises-Fisher distributions: and a fast implementation of $I_s(x)$. *Computational Statistics*, pages 1–14, February 2011.

[50] Rolf Sundberg. Maximum likelihood theory for incomplete data from an exponential family. *Scandinavian Journal of Statistics*, 1:49–58, 1974.

[51] Marc Teboulle. A unified continuous optimization framework for center-based clustering methods. *Journal of Machine Learning Research*, 8:65–102, 2007.

[52] Matus Telgarsky and Andrea Vattani. Hartigan's method: k-means clustering without Voronoi. *Journal of Machine Learning Research*, 9:820–827, 2010.

[53] Andrea Vattani. k-means requires exponentially many iterations even in the plane. *Discrete & Computational Geometry*, 45(4):596–616, 2011.